REVIEW

Bioinformatics applied to gene transcription regulation

Gioia Altobelli

Department of Endocrinology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

(Correspondence should be addressed to G Altobelli; Email: g.altobelli@qmul.ac.uk)

Abstract

Understanding regulation of gene transcription is central to molecular biology as well as being of great interest in medicine. The molecular syntax of the concerted transcriptional activation/repression of gene networks in mammal cells, which shape the physiological response to the molecular signals, is often unknown or not completely understood. Combining genome-wide experiments with *in silico* approaches opens the way to a more systematic comprehension of the molecular mechanisms of transcription regulation. Diverse bioinformatics tools have been developed to help unravel these mechanisms, by handling and processing data at different stages: from data collection and storage to the identification of molecular targets and from the detection of DNA motif signatures in the regulatory sequences of functionally related genes to the identification of relevant regulatory networks. Moreover, the large amount of genome-wide scale data recently produced has attracted professionals from diverse backgrounds to this cutting-edge realm of molecular biology. This mini-review is intended as an orientation for multidisciplinary professionals, introducing a streamlined workflow in gene transcription regulation with emphasis on sequence analysis. It provides an outlook on tools and methods, selected from a host of bioinformatics resources available today. It has been designed for the benefit of students, investigators, and professionals who seek a coherent yet quick introduction to *in silico* approaches to analyzing regulation of gene transcription in the post-genomic era.

Journal of Molecular Endocrinology (2012) **49**, R51–R59

Introduction

Completion of the human genome reference and development of next-generation sequencing (NGS) has led to an exciting era in which the complexity of biomedical research can be tackled in a new way, requiring multidisciplinary expertise at multiple levels (Noble 2002, Quackenbush 2006b, Navlakha & Bar-Joseph 2011). Bioinformatics, as an integrative discipline in this field, increasingly occupies a larger portion of the biomedical stage. Formerly seen as a service in wet laboratories or as a specialization in computer science departments, it is presently an intrinsic component of scientific discovery (Noble 2002). Significant benefits can derive from both sharing experience in a multidisciplinary team and employing individuals who integrate interdisciplinary expertise. Thus, it is important to prepare a new generation of biologists to master computational and quantitative skills (Bialek & Botstein 2004, Pevzner 2004, Wingreen & Botstein 2006, Dudley & Butte 2009, Pevzner & Shamir 2009) as well as knowledge in specialized areas of biomedicine. Moreover, creating and maintaining appropriate infrastructures helps in unifying the

heterogeneity of post-genomic data and dedicated bioinformatics toolkits. Large resources for data handling and manipulation, for example, open interactive systems such as Galaxy (Giardine *et al.* 2005), BioGPS (Wu *et al.* 2009), and the open source project R-Bioconductor (Gentleman *et al.* 2004), are highly welcome and point to a successful, cooperative model. However, most efforts cannot yet fully rely on such coherent supporting frameworks. Instead, expertise and knowledge of heterogeneous tools are required (Bialek & Botstein 2004, Pevzner 2004, Wasserman & Sandelin 2004, Quackenbush 2006b, Wingreen & Botstein 2006, Dudley & Butte 2009, Pevzner & Shamir 2009). Here, I want to collect a few reference landmarks for technology professionals who want to enter the post-genomic biology era and for biomedical investigators and students who want to build bioinformatics expertise on their own. Given the vastness of the subject, I won't be able to cover all areas of interest, nor provide deep insight into the complex issues underlying tool conception and development. By focusing on a streamlined analysis workflow, I want to stimulate and support the reader's intellectual curiosity and desire to develop individualized skills.

Background

The physiological responses of cells to any molecular signal are shaped by the molecular syntax of a concerted transcriptional activation/repression of many genes, referred to as regulatory networks. The major molecular events that underlie gene transcription primarily include the recognition of specific DNA elements known as transcription factor binding sites (TFBSs) by transcription factor (TF) proteins, as well as the recognition of cofactors that specifically link these protein–DNA binding events to the transcriptional molecular machinery (Wasserman & Sandelin 2004, Juven-Gershon *et al.* 2008). Besides, micro-RNA circuitry and a variety of posttranslation modifications, including the topical epigenetic markers, further modulate gene transcription with either functional or dysfunctional effect (Bartel 2009, Fabbri & Calin 2010, Portela & Esteller 2010, Lister *et al.* 2011). The packaging of DNA sequences around histone proteins to form nucleosomes and chromatin fibers (Hayes & Hansen 2001), that is, the structural organization of chromatin in the cell nuclei, masks/reveals TFBSs under different physiological conditions (Li *et al.* 2007, Park 2009). Various epigenetic modifications further alter the chromatin conformation, likely combining with each other to generate chromatin states (Ernst & Kellis 2010). In addition to a first layer of signals constituted by TFBS motifs, we can therefore picture a second layer of signals, defined by CTCF-binding sites that indicate active/inactive DNA sequence domains, or chromatin boundaries (Jothi *et al.* 2008, Jeziorska *et al.* 2009, Phillips & Corces 2009, Botta *et al.* 2010, Tolstorukov *et al.* 2011).

The architecture of a typical gene promoter is not rigorously defined, although a functional regulatory sequence comprises core elements (core promoter) and several enhancer/silencer elements scattered at various distances from the annotated transcriptional start sites (Wasserman & Sandelin 2004, Carninci *et al.* 2005, Halfon 2006, Juven-Gershon *et al.* 2008, Dolfini *et al.* 2009), as shown in Fig. 1. Enhancers often lie several kilo bases from the core, as observed in regulation by nuclear receptors (Lin *et al.* 2007, Kuttippurathu *et al.* 2011, Navlakha & Bar-Joseph 2011), and may lie in intragenic regions (mainly introns) as well as in the upstream sequence of a gene. The ‘linear’ architecture of the promoter (Fig. 1) in fact reflects the tri-dimensional structure of chromatin packaging, the conformation of which is consistent with a dynamic fractal globule able to fold/unfold any genomic locus, as recently revealed by Hi-C technique and Monte Carlo simulations (Lieberman-Aiden *et al.* 2009). The organization of chromatin on a genome scale deeply affects gene transcription regulation (Li *et al.* 2007, Ernst *et al.* 2011, Tolstorukov *et al.* 2011,

Figure 1 Schematic view of a hypothetical eukaryotic gene regulatory region. Right hand arrows stand for the transcriptional start site (TSS). Geometric shapes represent various DNA elements or motifs: blue shapes, various transcription factor binding sites; red trapezoid, the proximal core. Black spots represent chromatin boundaries such as CTCF-binding motifs. In higher dimensions (2D, top), the distal elements (1D, bottom far left) may be located closer to the core promoter and are therefore able to physically interact with the general transcriptional machinery. Multiple transcripts of a gene may be found in databases. Multiple TSSs have been observed for specific genes involved in developmental functions. Depending on the repository/species, the TSS may be associated with just one canonical transcript.

Delest *et al.* 2012). Further advancements in our understanding of gene transcriptional regulation will derive from the exploitation of structural findings on chromosome territories (Fraser & Bickmore 2007, Lieberman-Aiden *et al.* 2009, Shaw 2010), TF–TFBS binding events (Honig & Rohs 2011, Nikolova *et al.* 2011, Delest *et al.* 2012), and from a systemic approach (He *et al.* 2009, Ashworth *et al.* 2011).

In the current working model of transcription, proximal/distal regulatory elements are arranged in units, called cis-regulatory modules (CRMs), which are organized in a somewhat complex way (Halfon 2006, Jeziorska *et al.* 2009). Among other possible arrangements, a typical CRM consists of dense clusters of similar signals (Halfon 2006), which is a key feature exploited by the statistical methods developed for their detection (see below). The identification of CRMs has been an object of intensive bioinformatics research (Wasserman & Sandelin 2004, Elnitski *et al.* 2006, van Nimwegen 2007, Rister & Desplan 2010, Medina-Rivera *et al.* 2011). For a long time, the detection of DNA elements on a genomic scale was only attempted *in silico* and with limited success (Wasserman & Sandelin 2004, Hannonhalli 2008). The development of new technologies has promoted a systematic identification of TFBSs at high resolution. The cell and tissue specificity of the physiological response to molecular signals, for years addressed *in vitro* on a gene-by-gene basis, has been tackled *in vivo* via genome-wide approaches such as ChIP-on-Chip (Elnitski *et al.* 2006, Tavera-Mendoza *et al.* 2006, Dufour *et al.* 2007, Zheng *et al.* 2007) and ChIP-seq (Johnson *et al.* 2007, Barski & Zhao 2009, Park 2009, Pepke *et al.* 2009, Visel *et al.* 2009). In particular,

ChIP-seq technology has enabled the unbiased identification of all DNA sequences (TFBS) bound by a specific protein of interest (TF) (Park 2009, Liu *et al.* 2010) and is also successfully used in epigenomics (Bock & Lengauer 2008, Park 2008, Huss 2010, Lim *et al.* 2010, Ongenaert 2010). *In silico* approaches are currently applied in combination with genome-wide experimental techniques (D'Haeseleer 2006a, Elnitski *et al.* 2006, Dufour *et al.* 2007, Bickel *et al.* 2009, Gazdag *et al.* 2009, Delacroix *et al.* 2010).

The methods developed for detection of DNA motifs in the regulatory sequences of genes were devised in the framework of information theory (Hertz *et al.* 1990, Hertz & Stormo 1999), at a time when no structural clue about chromatin packaging was available. In this theory, the DNA molecule is simply a long word composed of four letters, the A G T C nucleotides, where the motif is hidden (Hertz *et al.* 1990, Hertz & Stormo 1999, Stormo 2000, Wasserman & Sandelin 2004, D'Haeseleer 2006a,b). The characterization of the (non-coding) DNA regions in terms of functional CRMs remains a computational challenge, especially for compositional signals, involving the combinatorial activation of different species of TFBSs (Bluthgen *et al.* 2005, Pierstorff *et al.* 2006, Van Loo & Marynen 2009). A comprehensive probabilistic framework for DNA motifs discovery, devised by van Nimwegen (2007), provides a unifying view of the sequence analysis approach to gene transcription regulation. The *in silico* determination of chromatin boundaries has been attempted from experimental chromatin signatures (Won *et al.* 2008) and nucleosome positioning prediction (Ioshikhes *et al.* 2006, Segal *et al.* 2006). The first systematic characterizations of chromatin states based on the *in silico* integration of a variety of epigenetic marks in human T-cells and *Drosophila* have been recently provided by Ernst & Kellis (2010), Ernst *et al.* (2011) and Riddle *et al.* (2011).

The presence of repeated, similar signals in the same non-coding DNA region, that is, the statistical overrepresentation of a DNA motif, especially when the region and/or the signal are conserved across the species, is key to inferring a biological role for this DNA element. This is a key criterion for detecting functional motifs, referred to as phylogenetic footprinting (Wasserman & Sandelin 2004, D'Haeseleer 2006a,b, Hannenhalli 2008) and implemented in several dedicated bioinformatics tools (see below). Moreover, statistical overrepresentation of (combinations of) DNA motifs in many of the regulatory regions of a set of co-expressed genes may indicate that these gene targets are co-regulated. In this set of regulatory regions, specific combinations of DNA motifs, henceforth 'DNA signatures', will be detected with good statistics compared with backgrounds (see Fig. 2).

Figure 2 A set of hypothetical regulatory regions for genes that are functionally related; for example, they may be downregulated by a known transcription factor and therefore called co-regulated (left); and a set of regulatory regions for genes that are selected at random, called background (right). Rectangles of different colors represent different TFBSs. In sequence analysis, the occurrence of DNA motifs in regulatory regions of functionally related genes is statistically compared to the occurrence of motifs in a background cohort. In this example, the count of orange motifs in co-regulated (left) exceeds the relevant count in background (right). A combination of green and orange motifs is also observed in co-regulated (all sequences but one bear it) and might be functionally important too.

Biochemical pathways target different DNA sequence elements and implement the gene expression control that is key to the regulatory events. In principle, knowledge of the distribution of the functional DNA elements in the regulatory regions of a set of responsive genes should enable the inference of the relevant regulatory networks. Inference of these networks from expression microarray data can be attempted by reverse engineering (He *et al.* 2009, Ashworth *et al.* 2011). Comprehensive coverage of reverse engineering and systemic approaches to gene transcription (regulatory) networks can be found in He *et al.* (2009), Koyuturk (2010) and Ashworth *et al.* (2011), and references therein. Much more difficult, and as yet unproven, is a forward approach, i.e. inferring the regulatory networks from the distribution of (active) DNA motifs in the gene regulatory regions (Altobelli 2007). In general, reconstruction of gene networks is bound to provide a static picture of the actual biological pathways. Time evolution must be addressed with an appropriate experimental design.

Bioinformatics toolkits

We distinguish here two levels of data analysis: 1) the treatment of raw data, which leads to the identification of target genes, and the unraveling of the biological meaning (ontological or functional analysis) of these targets; and 2) the integration of different types of information into a unifying view, the reconstruction of gene networks and biological pathways, and the

generation of regulative hypotheses. Sequence analysis may be performed at level 1, helping in identifying TBFSs from ChIP-seq data, for example, and/or at level 2 as in the case study below. In the following, we illustrate analysis tools for a hypothetical workflow that includes whole-genome techniques (expression microarray and ChIP-seq) and provide a concise applicative example of sequence analysis of regulatory regions of these gene targets. Even though many computational tools are available, the assumptions under which they have been developed may not fully apply to one's specific purpose. Thus, the analytical strategy is in best practice tailored for the specific biological system and research goals, possibly by controlling experiment design and always by evaluating the strengths and limitations of the method used. Both microarray and NGS data can be analyzed using open software packages within the R-Bioconductor framework (Gentleman *et al.* 2004) and/or Galaxy (Giardine *et al.* 2005). Methods and strategies for ChIP-seq analysis can be found in Pepke *et al.* (2009). Zhang *et al.* (2011) provide an extensive review on the impact of NGS technologies in genomics and Marra *et al.* in functional genomics and epigenomics (Morozova & Marra 2008, Hirst & Marra 2010). For foundations of gene expression analysis, see for example Quackenbush (2001, 2002, 2003 and 2006a) and Hsiao *et al.* (2005).

Once a list of (possibly) co-regulated genes is available, the detection of DNA signatures (Fig. 2) may be attempted using more than one method from two classes: 1) the algorithms that search for known motifs by screening the sequence against collections of TFBS models and 2) the algorithms that discover DNA motifs by unraveling the statistical property of the sequence, referred to as *de novo*. TBFSs are usually modeled by either consensus (Xie *et al.* 2005, D'Haeseleer 2006b) or matrix (Wasserman & Sandelin 2004, D'Haeseleer 2006a,b). The model-based methods are somewhat biased by the quality and type of the model used, yet provide a direct estimate of which TF is involved in regulation (Wasserman & Sandelin 2004). The *de novo* methods may spot novel motifs, yet require output interpretation; and both classes demand an appropriate definition of background in order to assign statistical significance (usually *P* values) to their outputs. A background set of sequences is usually constituted by the entire set of promoters of all genes for the species of interest or, for example, by the regulatory sequences of the genes that are functionally opposite, i.e. downregulated vs upregulated, depending on the problem under investigation.

An excellent web toolkit for (proximal) regulatory region analysis is the matrix-based Pscan (Zambelli *et al.* 2009). It is both user friendly and statistically robust, enabling fast detection of DNA signatures for several different promoter sizes and for several species.

Moreover, the source code can be downloaded for stand-alone usage. Tools for *de novo* motif discovery, that is from the second class of algorithms mentioned earlier, are often used in analysis of ChIP-seq experiments: WEEDER (Pavesi *et al.* 2004, Tompa *et al.* 2005), MEME (Bailey & Elkan 1994), MDScan (Liu *et al.* 2002), and W-ChIPMotifs (Jin *et al.* 2009) are equipped with web interfaces. Convenient integrative frames for launching multiple DNA discovery algorithms (two or more of the above-mentioned) are TAMO (Gordon *et al.* 2005), by shell scripting, and the web-based CompleteMOTIFs (Kuttiappurathu *et al.* 2011). Motif discovery algorithms eventually need to be coupled to other tools, either to link the discovered motifs to known TFBSs or to classify them using, for example, clustering algorithms (Bickel *et al.* 2009, Hackenberg *et al.* 2011). STAMP (Mahony & Benos 2007), a user-friendly tool that may serve this purpose, is also handled by the latest version of TAMO (Gordon *et al.* 2005). The more recent Tmod tool kit (Sun *et al.* 2010) allows the integration of 12 different motif discovery algorithms for Windows operating systems. RSAT (Thomas-Chollier *et al.* 2008) is a comprehensive interactive platform that supports hundreds of genomes and implements a variety of approaches to regulatory sequence analysis including unique visualization features. Based on multispecies alignments, OPOSSUM (Ho Sui *et al.* 2005) and CORG (Dieterich *et al.* 2005) are the major resources for investigating distal motifs, providing both collections of conserved regulatory regions and conserved motifs. The method implemented in REDUCE Suite v2 algorithm (Foat *et al.* 2006) instead models the interaction between TF and TFBS, enabling the inference of the TF sequence-specific binding affinity from a single expression experiment.

Gene annotation enrichment tools help in identifying the relevant (altered) biological processes in genome-wide experiments. The 68 gene enrichment tools developed for the functional analysis of gene expression data are compared by Huang *da et al.* (2009a,b). Enrichment tools heavily rely on the quality of annotation databases, which are incomplete and biased by construction (Huang *da et al.* 2009a, Liberzon *et al.* 2011). Comprehensive, web-based platforms are DAVID (Huang *da et al.* 2009b), GSEA (Subramanian *et al.* 2005, Liberzon *et al.* 2011), also available as an R-Bioconductor package and as a desktop application, and GREAT (McLean *et al.* 2010), which has been developed for unbiased functional annotation of genomic regions in human and mouse genomes. The latter is therefore recommended for annotating ChIP-seq-derived targets. For an extensive catalog of bioinformatics resources and web tools, the reader is referred to the dedicated annual publication by Brazas *et al.* (2011).

Figure 3 A sequence analysis workflow conceived for the characterization of regulatory regions of estrogen-responsive genes in breast cancer cells. In the blue rectangle: two algorithms for DNA sequence analysis of proximal regions (≤ 1 kbp) in combination with a collection of conserved motifs derived from a method enabling investigation of distal (up to 15 kbp) conserved nucleotide blocks. The input, which comprises several lists of estrogen-responsive genes and their relevant regulatory regions, was taken from a manually curated, in-house database (based on published experiments up to year 2005). The multiple outputs of this workflow were 1) sub-lists of genes associated by the presence of (at least) one motif/TBFS, 2) overrepresented motifs and their genomic localizations, 3) lists of putative transcription factors, and, after data integration, 4) regulatory hypotheses.

Case study

In order to tackle the complexity of estrogenic regulation in breast cancer cells *in silico*, a large sequence analysis workflow was set up as in Fig. 3 (Altobelli 2007). Lists of estrogen target genes were collected and organized in a manually curated, in-house database, by selecting those that exhibited an early response to estradiol in breast cancer cells (up/downregulation within 4 h), from gene expression microarray and gene-by-gene experiments published up to year 2005. The workflow combined approaches for DNA sequence analysis of proximal regions (≤ 1 kbp) with a method that enabled the investigation of distal (up to 15 kbp) conserved nucleotide blocks (Altobelli 2007). In particular, the above-mentioned Pscan (Zambelli *et al.* 2009) and WEEDER (Pavesi *et al.* 2004, Tompa *et al.* 2005) algorithms were used, as well as a collection of conserved motifs generated from a modified version of CORG (Dieterich *et al.* 2003). One of the many purposes of this study was to formulate regulatory hypotheses to be tested in the laboratory. Estrogen-responsive elements (EREs) are highly degenerate (Carroll & Brown 2006) and found scattered over large genomic distances (Carroll & Brown 2006). Thus, they are very elusive to detection *in silico* (Sismond *et al.* 2007). Estrogenic regulation often occurs through the interaction of (activated) estrogen

receptor (ER) with cognate TFs, exhibiting TFBSs adjacent to EREs (co-localized TFBSs). This suggests that DNA elements of different types located within a distance comparable to the nucleosome average size (circa 200 bp) may form a functional unit. One of the implemented strategies, therefore, was searching for EREs in the surroundings of other TFBSs – the latter detected with robust statistics.

The workflow provided multiple outputs as follows: 1) sub-lists of genes associated by the presence of at least one DNA motif, 2) DNA motifs and their genomic localizations (DNA signatures), 3) lists of putative TFs, and (after literature mining and data integration) 4) regulatory hypotheses. Among other DNA signatures, several conserved GATA family motifs were detected in multiple occurrences in the upstream region of a small group of developmental genes including cyclin-G2 (CCNG2). Two (novel) degenerated ERE motifs were also detected in the proximity of two conserved GATA3 binding sites in the CCNG2 upstream regulatory region (Altobelli 2007). Cyclin-G2, a cell cycle inhibitor, is strongly downregulated in breast cancer cells and is a primary ER target gene (Stossi *et al.* 2006). The co-localized ERE and GATA motifs suggest that an interaction between the ER and GATA3 may be implicated in the regulation of the above-mentioned group of developmental genes and that this might play a role in mammary gland development and cancer

differentiation (Chou *et al.* 2010). This interaction might, in fact, contribute to the shaping of the lactiferous ducts, determining the fate of luminal cells, which express estrogen receptor (ER+), compared to myo-epithelial cells, which do not express estrogen receptor (ER−). This may also have implications in the characterization of types of breast cancer (ER+ vs ER− luminal-like types).

Genome-wide experiments in estrogen-responsive cells have recently enabled the detection of presumably direct targets of ERs, as well as a catalog of EREs and binding sites for putative cognate TFs (Carroll *et al.* 2005, Carroll & Brown 2006, Lin *et al.* 2007). This makes it possible, in principle, to validate all *in silico* outcomes and to refine the workflow accordingly.

Declaration of interest

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the review.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Acknowledgements

The author thanks all the students and colleagues whose persistent questions eventually prompted to write this work as well as Prof. R Bargar, Prof. R Flower, and Mr J Maftin for kindly proofreading the manuscript.

References

- Altobelli G 2007 Functional classification of estrogen-responsive gene regulatory sequences in breast cancer cells: towards the identification of regulatory networks. *PhD Thesis*. University of Turin http://dott-scsv.campusnet.unito.it/do/home.pl/View?doc=archivio_tesi.html.
- Ashworth J, Wurtmann EJ & Baliga NS 2011 Reverse engineering systems models of regulation: discovery, prediction and mechanisms. *Current Opinion in Biotechnology*. <http://dx.doi.org/10.1016/j.cbb.2011.03.031>.
- Bailey TL & Elkan C 1994 Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* **2** 28–36.
- Barski A & Zhao K 2009 Genomic location analysis by ChIP-Seq. *Journal of Cellular Biochemistry* **107** 11–18. (doi:10.1002/jcb.22077)
- Bartel DP 2009 MicroRNAs: target recognition and regulatory functions. *Cell* **136** 215–233. (doi:10.1016/j.cell.2009.01.002)
- Bialek W & Botstein D 2004 Introductory science and mathematics education for 21st-Century biologists. *Science* **303** 788–790. (doi:10.1126/science.1095480)
- Bickel PJ, Brown JB, Huang H & Li Q 2009 An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **367** 4313–4337. (doi:10.1098/rsta.2009.0164)
- Bluthgen N, Kielbasa SM & Herzel H 2005 Inferring combinatorial regulation of transcription *in silico*. *Nucleic Acids Research* **33** 272–279. (doi:10.1093/nar/gki167)
- Bock C & Lengauer T 2008 Computational epigenetics. *Bioinformatics* **24** 1–10. (doi:10.1093/bioinformatics/btm546)
- Botta M, Haider S, Leung IX, Lio P & Mozziconacci J 2010 Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Molecular Systems Biology* **6** 426. (doi:10.1038/msb.2010.79)
- Brazas MD, Yim DS, Yamada JT & Ouellette BF 2011 The 2011 Bioinformatics Links Directory update: more resources, tools and databases and features to empower the bioinformatics community. *Nucleic Acids Research* **39** W3–W7. (doi:10.1093/nar/gkr514)
- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al.* 2005 The transcriptional landscape of the mammalian genome. *Science* **309** 1559–1563. (doi:10.1126/science.1112014)
- Carroll JS & Brown M 2006 Estrogen receptor target gene: an evolving concept. *Molecular Endocrinology* **20** 1707–1714. (doi:10.1210/me.2005-0334)
- Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoutte J, Shao W, Hestermann EV, Geistlinger TR *et al.* 2005 Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122** 33–43. (doi:10.1016/j.cell.2005.05.008)
- Chou J, Provost S & Werb Z 2010 GATA3 in development and cancer differentiation: cells GATA have it!. *Journal of Cellular Physiology* **222** 42–49. (doi:10.1002/jcp.21943)
- Delacroix L, Moutier E, Altobelli G, Legras S, Poch O, Choukrallah MA, Bertin I, Jost B & Davidson I 2010 Cell-specific interaction of retinoic acid receptors with target genes in mouse embryonic fibroblasts and embryonic stem cells. *Molecular and Cellular Biology* **30** 231–244. (doi:10.1128/MCB.00756-09)
- Delest A, Sexton T & Cavalli G 2012 Polycomb: a paradigm for genome organization from one to three dimensions. *Current Opinion in Cell Biology* **24** 405–414. (doi:10.1016/j.ccb.2012.01.008)
- D'Haeseleer P 2006a How does DNA sequence motif discovery work? *Nature Biotechnology* **24** 959–961. (doi:10.1038/nbt0806-959)
- D'Haeseleer P 2006b What are DNA sequence motifs? *Nature Biotechnology* **24** 423–425. (doi:10.1038/nbt0406-423)
- Dieterich C, Wang H, Rateitschak K, Luz H & Vingron M 2003 CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Research* **31** 55–57. (doi:10.1093/nar/gkg007)
- Dieterich C, Grossmann S, Tanzer A, Ropcke S, Arndt PF, Stadler PF & Vingron M 2005 Comparative promoter region analysis powered by CORG. *BMC Genomics* **6** 24. (doi:10.1186/1471-2164-6-24)
- Dolfini D, Zambelli F, Pavesi G & Mantovani R 2009 A perspective of promoter architecture from the CCAAT box. *Cell Cycle* **8** 4127–4137. (doi:10.4161/cc.8.24.10240)
- Dudley JT & Butte AJ 2009 A quick guide for developing effective bioinformatics programming skills. *PLoS Computational Biology* **5** e1000589. (doi:10.1371/journal.pcbi.1000589)
- Dufour CR, Wilson BJ, Huss JM, Kelly DP, Alaynick WA, Downes M, Evans RM, Blanchette M & Giguere V 2007 Genome-wide orchestration of cardiac functions by the orphan nuclear receptors ERR α and γ . *Cell Metabolism* **5** 345–356. (doi:10.1016/j.cmet.2007.03.007)
- Elnitski L, Jin VX, Farnham PJ & Jones SJ 2006 Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Research* **16** 1455–1464. (doi:10.1101/gr.4140006)
- Ernst J & Kellis M 2010 Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28** 817–825. (doi:10.1038/nbt.1662)

- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al.* 2011 Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473** 43–49. (doi:10.1038/nature09906)
- Fabbri M & Calin GA 2010 Epigenetics and miRNAs in human cancer. *Advances in Genetics* **70** 87–99. (doi:B978-0-12-380866-0.60004-6[pil])
- Foat BC, Morozov AV & Bussemaker HJ 2006 Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22** e141–e149. (doi:10.1093/bioinformatics/btl223)
- Fraser P & Bickmore W 2007 Nuclear organization of the genome and the potential for gene regulation. *Nature* **447** 413–417. (doi:10.1038/nature05916)
- Gazdag E, Santenard A, Ziegler-Birling C, Altobelli G, Poch O, Tora L & Torres-Padilla ME 2009 TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. *Genes and Development* **23** 2210–2223. (doi:10.1101/gad.535209)
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al.* 2004 Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5** R80. (doi:10.1186/gb-2004-5-10-r80)
- Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al.* 2005 Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* **15** 1451–1455. (doi:10.1101/gr.4086505)
- Gordon DB, Nekudova L, McCallum S & Fraenkel E 2005 TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics* **21** 3164–3165. (doi:10.1093/bioinformatics/bti481)
- Hackenberg M, Carpena P, Bernaola-Galvan P, Barturen G, Alganza AM & Oliver JL 2011 WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms for Molecular Biology* **6** 2. (doi:10.1186/1748-7188-6-2)
- Halfon MS 2006 (Re)modeling the transcriptional enhancer. *Nature Genetics* **38** 1102–1103. (doi:10.1038/ng1006-1102)
- Hannenhalli S 2008 Eukaryotic transcription factor binding sites – modeling and integrative search methods. *Bioinformatics* **24** 1325–1331. (doi:10.1093/bioinformatics/btn198)
- Hayes JJ & Hansen JC 2001 Nucleosomes and the chromatin fiber. *Current Opinion in Genetics & Development* **11** 124–129. (doi:10.1016/S0959-437X(00)00168-4)
- He F, Balling R & Zeng AP 2009 Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. *Journal of Biotechnology* **144** 190–203. (doi:10.1016/j.jbiotec.2009.07.013)
- Hertz GZ & Stormo GD 1999 Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** 563–577. (doi:10.1093/bioinformatics/15.7.563)
- Hertz GZ, Hartzell GW III & Stormo GD 1990 Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences* **6** 81–92.
- Hirst M & Marra MA 2010 Next generation sequencing based approaches to epigenomics. *Briefings in Functional Genomics* **9** 455–465. (doi:10.1093/bfpg/elq035)
- Honig B & Rohs B 2011 Flipping Watson and Crick. *Nature* **470** 472–473. (doi:10.1038/470472a)
- Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP & Wasserman WW 2005 oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Research* **33** 3154–3164. (doi:10.1093/nar/gki624)
- Hsiao A, Ideker T, Olefsky JM & Subramaniam S 2005 VAMPIRE microarray suite: a web-based platform for the interpretation of gene expression data. *Nucleic Acids Research* **33** W627–W632. (doi:10.1093/nar/gki443)
- Huang da W, Sherman BT & Lempicki RA 2009a Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37** 1–13. (doi:10.1093/nar/gkn923)
- Huang da W, Sherman BT & Lempicki RA 2009b Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4** 44–57. (doi:10.1038/nprot.2008.211)
- Huss M 2010 Introduction into the analysis of high-throughput-sequencing based epigenome data. *Briefings in Bioinformatics* **11** 512–523. (doi:10.1093/bib/bbq014)
- Ioshikhes IP, Albert I, Zanton SJ & Pugh BF 2006 Nucleosome positions predicted through comparative genomics. *Nature Genetics* **38** 1210–1215. (doi:10.1038/ng1878)
- Jeziorska DM, Jordan KW & Vance KW 2009 A systems biology approach to understanding cis-regulatory module function. *Seminars in Cell & Developmental Biology* **20** 856–862. (doi:10.1016/j.semcdb.2009.07.007)
- Jin VX, Apostolos J, Nagisetty NS & Farnham PJ 2009 W-ChIPMotifs: a web application tool for *de novo* motif discovery from ChIP-based high-throughput data. *Bioinformatics* **25** 3191–3193. (doi:10.1093/bioinformatics/btp570)
- Johnson DS, Mortazavi A, Myers RM & Wold B 2007 Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316** 1497–1502. (doi:10.1126/science.1141319)
- Jothi R, Cuddapah S, Barski A, Cui K & Zhao K 2008 Genome-wide identification of *in vivo* protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Research* **36** 5221–5231. (doi:10.1093/nar/gkn488)
- Juven-Gershon T, Hsu JY, Theisen JW & Kadonaga JT 2008 The RNA polymerase II core promoter – the gateway to transcription. *Current Opinion in Cell Biology* **20** 253–259. (doi:10.1016/j.ceb.2008.03.003)
- Koyuturk M 2010 Algorithmic and analytical methods in network biology. *Wiley Interdisciplinary Reviews Systems Biology and Medicine* **2** 277–292. (doi:10.1002/wsbm.61)
- Kuttipurathu L, Hsing M, Liu Y, Schmidt B, Maskell DL, Lee K, He A, Pu WT & Kong SW 2011 CompleteMOTIFs: DNA, motif discovery platform for transcription factor binding experiments. *Bioinformatics* **27** 715–717. (doi:10.1093/bioinformatics/btq707)
- Li B, Carey M & Workman JL 2007 The role of chromatin during transcription. *Cell* **128** 707–719. (doi:10.1016/j.cell.2007.01.015)
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P & Mesirov JP 2011 Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27** 1739–1740. (doi:10.1093/bioinformatics/btr260)
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.* 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326** 289–293. (doi:10.1126/science.1181369)
- Lim SJ, Tan TW & Tong JC 2010 Computational epigenetics: the new scientific paradigm. *Bioinformatics* **4** 331–337. (doi:10.6026/97320630004331)
- Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F *et al.* 2007 Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genetics* **3** e87. (doi:10.1371/journal.pgen.0030087)
- Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S *et al.* 2011 Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471** 68–73. (doi:10.1038/nature09798)
- Liu XS, Brutlag DL & Liu JS 2002 An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* **20** 835–839. (doi:10.1038/nbt717)

- Liu ET, Pott S & Huss M 2010 Q&A: ChIP-seq technologies and the study of gene regulation. *BMC Biology* **8** 56. (doi:10.1186/1741-7007-8-56)
- Mahony S & Benos PV 2007 STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research* **35** W253–W258. (doi:10.1093/nar/gkm272)
- McLean CX, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM & Bejerano G 2010 GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28** 495–501. (doi:10.1038/nbt.1630)
- Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J & van Helden J 2011 Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research* **39** 808–824. (doi:10.1093/nar/gkq710)
- Morozova O & Marra MA 2008 Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92** 255–264. (doi:10.1016/j.ygeno.2008.07.001)
- Navlakha S & Bar-Joseph Z 2011 Algorithms in nature: the convergence of systems biology and computational thinking. *Molecular Systems Biology* **7** 546. (doi:10.1038/msb.2011.78)
- Nikolova EN, Kim E, Wise AA, O'Brien PJ, Andricioaei I & Al-Hashimi HM 2011 Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470** 498–502. (doi:10.1038/nature09775)
- van Nimwegen E 2007 Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics* **8** (Suppl 6) S4. (doi:10.1186/1471-2105-8-S6-S4)
- Noble D 2002 The rise of computational biology. *Nature Reviews Molecular and Cellular Biology* **3** 459–463. (doi:10.1038/nrm810)
- Ongenaert M 2010 Epigenetic databases and computational methodologies in the analysis of epigenetic datasets. *Advances in Genetics* **71** 259–295. (doi:10.1016/B978-0-12-380864-6.00009-2)
- Park PJ 2008 Epigenetics meets next-generation sequencing. *Epigenetics* **3** 318–321. (doi:10.4161/epi.3.6.7249)
- Park PJ 2009 ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10** 669–680. (doi:10.1038/nrg2641)
- Pavesi G, Merighetti P, Mauri G & Pesole G 2004 Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32** W199–W203. (doi:10.1093/nar/gkh465)
- Pepke S, Wold B & Mortazavi A 2009 Computation for ChIP-seq and RNA-seq studies. *Nature Methods* **6** S22–S32. (doi:10.1038/nmeth.1371)
- Pevzner PA 2004 Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians. *Bioinformatics* **20** 2159–2161. (doi:10.1093/bioinformatics/bth217)
- Pevzner P & Shamir R 2009 Computing has changed biology—biology education must catch up. *Science* **325** 541–542. (doi:10.1126/science.1173876)
- Phillips JE & Corces VG 2009 CTCF: master weaver of the genome. *Cell* **137** 1194–1211. (doi:10.1016/j.cell.2009.06.001)
- Pierstorff N, Bergman CM & Wiehe T 2006 Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics* **22** 2858–2864. (doi:10.1093/bioinformatics/btl499)
- Portela A & Esteller M 2010 Epigenetic modifications and human disease. *Nature Biotechnology* **28** 1057–1068. (doi:10.1038/nbt.1685)
- Quackenbush J 2001 Computational analysis of microarray data. *Nature Reviews Genetics* **2** 418–427. (doi:10.1038/35076576)
- Quackenbush J 2002 Microarray data normalization and transformation. *Nature Genetics* **32** (Suppl) 496–501. (doi:10.1038/ng1032)
- Quackenbush J 2003 Genomics. Microarrays – guilt by association. *Science* **302** 240–241. (doi:10.1126/science.1090887)
- Quackenbush J 2006a Computational approaches to analysis of DNA microarray data. *Yearbook of Medical Informatics* 91–103.
- Quackenbush J 2006b From 'omes to biology. *Animal Genetics* **37** (Suppl 1) 48–56. (doi:10.1111/j.1365-2052.2006.01476.x)
- Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D *et al.* 2011 Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Research* **21** 147–163. (doi:10.1101/gr.110098.110)
- Rister J & Desplan C 2010 Deciphering the genome's regulatory code: the many languages of DNA. *BioEssays* **32** 381–384. (doi:10.1002/bies.200900197)
- Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP & Widom J 2006 A genomic code for nucleosome positioning. *Nature* **442** 772–778. (doi:10.1038/nature04979)
- Shaw PJ 2010 Mapping chromatin conformation. *F1000 Biology Reports* **2** 18. (doi:10.3410/B2-18)
- Sismondi P, Biglia N, Ponzone R, Fusco L, Scafoglio C, Cicatiello L, Ravo M, Weisz A, Cimino D, Altobelli G *et al.* 2007 Influence of estrogens and antiestrogens on the expression of selected hormone-responsive genes. *Maturitas* **57** 50–55. (doi:10.1016/j.maturitas.2007.02.019)
- Stormo GD 2000 DNA binding sites: representation and discovery. *Bioinformatics* **16** 16–23. (doi:10.1093/bioinformatics/16.1.16)
- Stossi F, Likhite VS, Katzenellenbogen JA & Katzenellenbogen BS 2006 Estrogen-occupied estrogen receptor represses cyclin G2 gene expression and recruits a repressor complex at the cyclin G2 promoter. *Journal of Biological Chemistry* **281** 16272–16278. (doi:10.1074/jbc.M513405200)
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al.* 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102** 15545–15550. (doi:10.1073/pnas.0506580102)
- Sun H, Yuan Y, Wu Y, Liu H, Liu JS & Xie H 2010 Tmod: toolbox of motif discovery. *Bioinformatics* **26** 405–407. (doi:10.1093/bioinformatics/btp681)
- Tavera-Mendoza LE, Mader S & White JH 2006 Genome-wide approaches for identification of nuclear receptor target genes. *Nuclear Receptor Signaling* **4** e018.
- Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohee S & van Helden J 2008 RSAT: regulatory sequence analysis tools. *Nucleic Acids Research* **36** W119–W127. (doi:10.1093/nar/gkn304)
- Tolstorukov MY, Volfovsky N, Stephens RM & Park PJ 2011 Impact of chromatin structure on sequence variability in the human genome. *Nature Structural & Molecular Biology* **18** 510–515. (doi:10.1038/nsmb.2012)
- Tomba M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al.* 2005 Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* **23** 137–144. (doi:10.1038/nbt1053)
- Van Loo P & Marynen P 2009 Computational methods for the detection of cis-regulatory modules. *Briefings in Bioinformatics* **10** 509–524. (doi:10.1093/bib/bbp025)
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F *et al.* 2009 ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457** 854–858. (doi:10.1038/nature07730)
- Wasserman WW & Sandelin A 2004 Applied bioinformatics for the identification of regulatory elements. *Nature Reviews. Genetics* **5** 276–287. (doi:10.1038/nrg1315)
- Wingreen N & Botstein D 2006 Back to the future: education for systems-level biologists. *Nature Reviews. Molecular and Cellular Biology* **7** 829–832. (doi:10.1038/nrm2023)
- Won KJ, Chepelev I, Ren B & Wang W 2008 Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* **9** 547. (doi:10.1186/1471-2105-9-547)
- Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, James J, Huss JW *et al.* 2009 BioGPS: an extensible and

- customizable portal for querying and organizing gene annotation resources. *Genome Biology* **10** R130. (doi:10.1186/gb-2009-10-11-r130)
- Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES & Kellis M 2005 Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434** 338–345. (doi:10.1038/nature03441)
- Zambelli F, Pesole G & Pavesi G 2009 Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Research* **37** W247–W252. (doi:10.1093/nar/gkp464)
- Zhang J, Chiodini R, Badr A & Zhang G 2011 The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38** 95–109. (doi:10.1016/j.jgg.2011.02.003)
- Zheng M, Barrera LO, Ren B & Wu YN 2007 ChIP-chip: data, model, and analysis. *Biometrics* **63** 787–796. (doi:10.1111/j.1541-0420.2007.00768.x)

Received in final form 15 June 2012

Accepted 21 June 2012

Made available online as an Accepted Preprint 21 June 2012